认知建模中模型比较的方法

郭鸣谦[1] 潘晚坷[2] 胡传鹏[2]

(1 Behavioral Science Institute Radboud University, Nijmegen 6525GD;
2 南京师范大学心理学院，南京 21002)

摘　要：认知建模近年来在科学心理学获得广泛应用，而模型比较是认知建模中关键的一环：研究者需要通过模型比较来选择出最优模型，才能进行后续的假设检验或潜变量推断。模型比较不仅要考虑模型对数据的拟合（平衡过拟合与欠拟合），也需要考虑参数数据和数学形式的复杂度。然而，模型比较指标众多，纷繁复杂。将认知建模常用的模型比较的指标分为三大类，并介绍了其计算方法及优劣，包括拟合优度指标（包括平均平方误差、决定系数、RUC 曲线等）、基于交叉验证的指标（包括 AIC、DIC 等）和基于边际似然的指标。结合正交 Go /No-Go 范式下的数据，展示各指标在 R 语言中如何实现。在此基础上，探讨各指标的适用情境，介绍模型平均等模型比较的新思路。

关键词：认知建模；计算模型；模型选择

通讯作者：胡传鹏，男，博士，南京师范大学心理学院教授，E-mail: hcp4715@hotmail.com

最近的二十年来，基于计算模型(Computational models)对行为数据进行认知建模(Cognitive modeling)的研究越来越多受到研究者的关注。例如，在感知觉决策(Perceptual decision-making)领域中的贝叶斯感知觉模型(Bayesian perception model)(Kording & Wolpert, 2006)和漂移扩散模型(Drift diffusion model)(Forstmann et al., 2016; Ratcliff et al., 2016)等在认知神经科学得到了广泛的应用。类似的，强化学习模型(Reinforcement learning model)在价值决策(Value-based decision-making)研究中日益成为主流，其通过模型估计出的隐变量“预期误差(Prediction error)”可以有效地预测学习过程中多巴胺神经元(dopaminergic neuron)的活动(Schultz et al., 1997; Steinberg et al., 2013)。计算模型也是计算精神病学(Computational psychiatry)这一新兴交叉领域的基础(Geng et al., 2022; Huys et al., 2016; Montague et al., 2012; 区健新, 2020)，增进理解精神疾病人群的认知加工上的缺陷以提高对精神疾病诊断和分类的准确度，提供精准治疗(Pedersen et al., 2021)。

认知模型的步骤大致包括模拟数据(Simulation)、参数估计(Parameter estimation)、模型比较(Model comparison)和隐变量推断(Latent variable inference)(Wilson & Collins, 2019)等步骤。具体而言，研究者根据不同理论提出相应的计算模型进行模拟，并设计实验收集数据，使用各个计算模型拟合数据，通过模型比较来选出最优模型，最后根据最优模型进一步分析数据，将模型的中的隐变量与神经数据结合进行推断。

模型比较是认知建模里至关重要的一环，它不仅在认知建模中使用，也是各种涉及到计算模型的场景中必不可少的步骤。然而，心理学/认知科学等领域研究者对于模型比较的过程较为陌生，面对种类繁多的模型比较指标时，常感到困惑。此外，当前文献中也缺乏对模型比较的诸多方法进行系统梳理。有鉴于此，本文梳理模型比较的原则和各个方法，帮助读者理解当前模型比较背后的原理和适用情境，推动更好地运用认知建模。虽然本文的重点放在实验心理学里的认知建模当中，但是介绍的指标也可以应用于其他心理学常见的统计模型，例如分层线性回归、结构方程模型等等。

我们将首先介绍模型比较的基本原则，随后结合案例系统地介绍常见模型比较指标的原理和优缺点，最后，从实际应用的角度，总结各个指标的优劣和使用注意事项。

1　模型比较的基本原则

对于研究者而言，一个好的模型必须要具备如下两点特质。第一，它能够很好地解释或者拟合当前样本数据的模型。第二，模型要具有泛化能力，即能够对于当前数据之外的

数据同样提供较好的解释（即预测能力）。如果某个模型无法准确地解释当前样本数据，则可认为这个模型是欠拟合的(Underfitting)。如果某个模型能够非常好地解释当前样本数据但无法解释样本外的数据时，则认为这个模型过拟合的(Overfitting)(Friedman et al., 2001)。

研究者通常使用泛化误差(Generalization error)，即模型预测和真实数据的差异来衡量模型的泛化能力。泛化误差可以被分为方差(Variance)、偏差(Bias)和误差项(Irreducible error)。偏差是模型预测和真实数据之间的差异，方差表示模型在不同训练数据上预测结果的变化程度。模型难以同时达到小的偏差和方差，概因样本数据中存在噪音，过于复杂的模型虽然对样本数据拟合很好(此时的偏差很小)，却会将过多噪音考虑在内，令模型的预测极为不稳定(方差很大)。因此，随着模型的复杂度的增大，模型的偏差会逐渐减小，方差则会增大，这被称作偏差-方差权衡(Bias-variance trade-off)。偏差大的模型欠拟合，而方差大的模型则过拟合(Friedman et al., 2001)。选择模型是一个权衡模型的偏差和方差，从而使得模型的泛化误差最小的过程。

虽然模型的复杂度对其泛化能力有着重要作用，但其也受到诸多因素的影响。Myung and Pitt (1997)总结三种影响模型复杂度的因素。第一是模型的参数数量。一般情况下模型的参数越多复杂度越高。第二是模型的数学形式。例如，非线性的模型要比线性模型更复杂。第三是模型的参数空间范围。更大的参数空间范围说明模型拥有更多的自由度，也意味着模型更复杂。

根据模型比较指标关注点和原理的差异，可将它们分为三类。第一类为模型拟合优度(Goodness of fit)，这一类指标并没有考虑模型的复杂度，只是单纯地衡量模型对于当前样本数据的拟合程度。第二类是交叉验证(Cross validation)以及近似交叉验证的指标，这类指标关注于模型的泛化能力(Generalization ability)，即基于当前样本数据拟合后的模型对于样本外数据预测准确度(Out of sample prediction accuracy)。第三类是基于边际似然的指标[1] $P(y|M)$，其中y表示观测数据，M表示模型。边际似然着重于选择出候选模型里可能存在的“真实模型”。后二者都具有在复杂度和拟合优度之间进行权衡的特质。不同的模型比较指标各有其优缺点，不存在某一个指标全面优于他者。因此，研究者需要根据实际情况选择合适的指标。以下将通过一个数据作为示例，分别介绍这三大类指标。

2 示例数据

[1] 在贝叶斯统计中，边际似然(Marginal likelihood)也称为称模型证据(Model evidence)。

本文将结合正交 Go /No Go 范式的示例实验来介绍各模型指标的计算方法及特点(Cavanagh et al., 2013; Dorfman & Gershman, 2019; Guitart-Masip et al., 2012)。示例所用数据为使用下文介绍的认知模型模拟产生。模拟数据和后续模型比较指标的计算使用了 R 语言，具体代码见在线材料： https://github.com/zaizibai/model_comparison。

正交 Go/No Go 范式常被用于研究巴浦洛夫学习和工具性学习之间的关系。该范式是 2 × 2 的被试内实验设计，其中第一个变量是反应刺激：Go 和 No Go；第二个变量是行为反应后的反馈类型：获得奖励和避免惩罚。反应刺激和反馈类型两个条件结合起来共形成四种实验条件(在该范式中被称作提示符号 cue)：Go-获得奖赏、Go-避免惩罚，No Go-获得奖赏，No Go-避免惩罚。值得注意的是，每个条件下的正负反馈都是概率的。例如在“Go-避免惩罚”条件下，正确反应(即 Go)有 80%概率避免惩罚，但仍有 20%概率被惩罚；而错误反应（即 No-Go）则有 80%概率被惩罚，20%概率避免惩罚。实验开始时，被试并不知道每类条件下正确的反应，需要根据反馈不断地来学习。根据学习理论，在该范式里当反馈是获得奖赏时，人们易有 Go 反应；当反馈是避免惩罚时，则更容易产生 No Go 反应(Dayan et al., 2006)。

研究者通常使用简单的强化学习模型对该范式下的数据进行建模。该模型认为人类决策受两种学习因素影响：巴浦洛夫学习和工具性学习。工具性学习源自斯金纳的工具性学习理论，是刺激-反应-结果(Stimulus-Response-Outcome, SRO)的联结，而巴浦洛夫学习则是刺激-结果的联结，与反应无关。具体而言，选择 Go 或 No Go 反应的决策权重的公式如下：

$$w = b + Q + \pi \times V \quad (1)$$

这其中b代表个体对 Go 或 No Go 反应的天然的偏好，Q是工具性学习的决策变量，而V则是巴浦洛夫效应的决策变量，π是它的度量参数。关于该模型的具体细节，可以详见 Betts et al. (2020)或 Swart et al. (2017)。

本文中我们将使用结合了巴浦洛夫效应和工具性学习的模型模拟 10 个被试的数据，并拟合两种模型，包括模拟数据的真实模型(模型一)，以及没有巴浦洛夫效应而只有工具性学习的模型(模型二)。具体的拟合中，将使用分层贝叶斯模型(Hierarchical Bayesian estimation, HBE)和最大化后验概率法(Maximum a posterior estimation, MAP)。在接下来的部分，本文将结合案例模型和数据，具体介绍一些指标的计算方式。

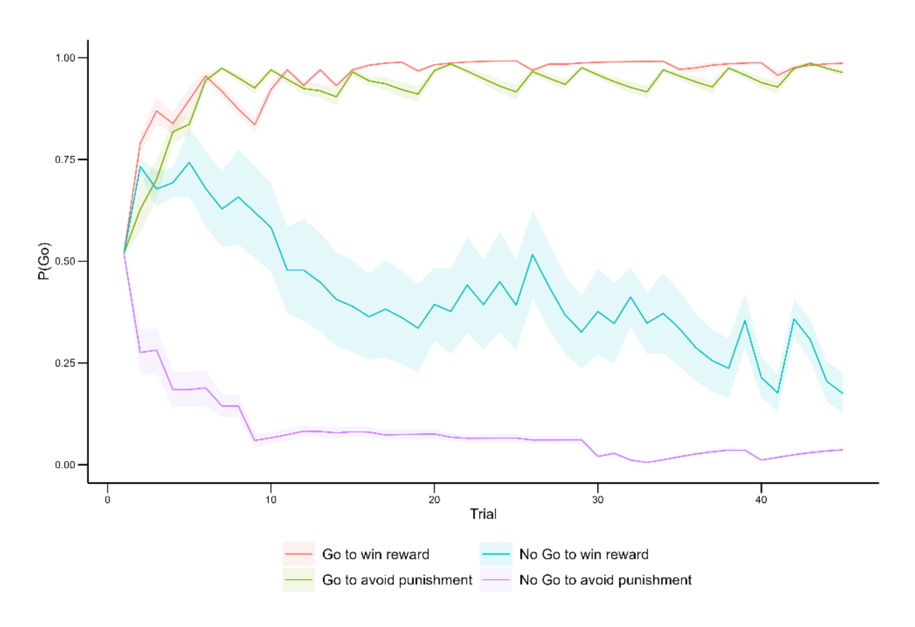


图 1. 案例 Trial-by-trial 的行为数据，由模型一生成。图中横坐标是试次数量，纵坐标是选择 Go 反应的比例。四种颜色代表了四种 cue。随着试次数量的增大，个体行为逐渐变得稳定，这体现了工具性学习的作用。而获得奖赏和避免惩罚 cue 下，个体 Go 反应的比例的不对称性则体现了巴浦洛夫效应。具体而言，个体更易有 Go 反应去获得奖赏，但是却更多地有 No Go 反应去避免惩罚。

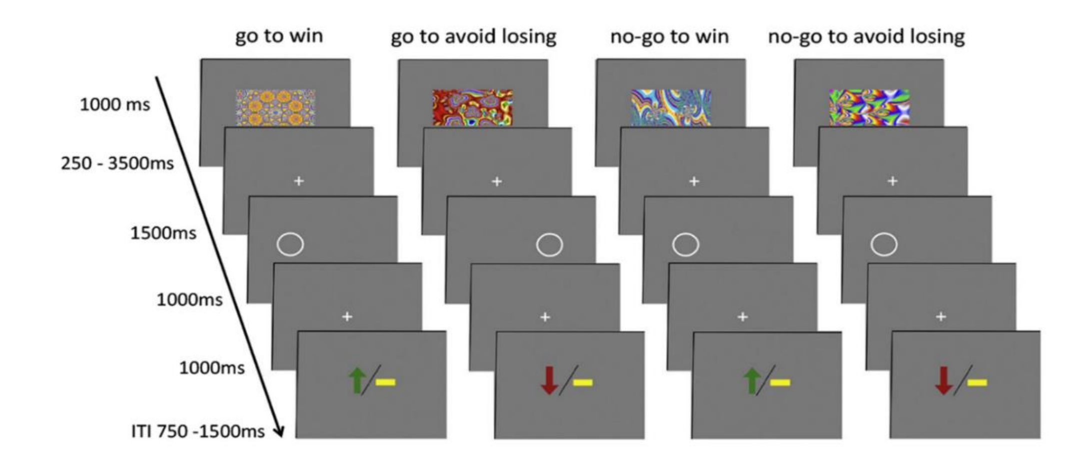


图 2. 案例的实验设计，引自 Betts et al. (2020)。单个试次的流程如下，被试首先会看到一个 cue，在 cue 消失后需进行 Go 或者 No Go 反应，反应完毕屏幕会呈现反应结果。在此任务里，被试需要去主动学习不同的 cue 的正确反应，以及正确结果是避免惩罚还是获得奖赏。



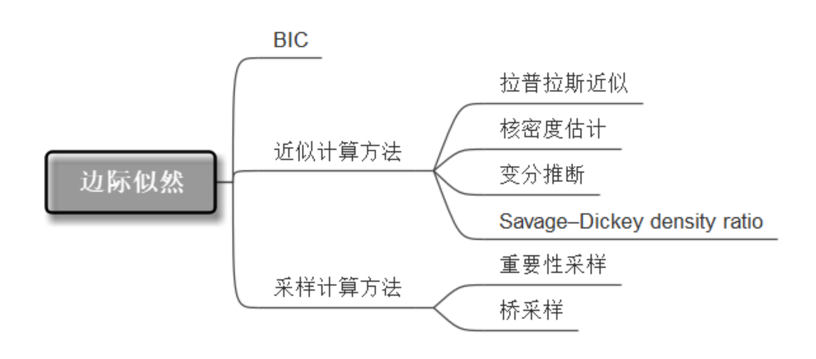


图 3. 认知建模里三种常见的模型比较指标，分别包括拟合优度指标、基于交叉验证的指标和基于边际似然的指标。

3 拟合优度指标

模型的拟合优度(Goodness of fit)主要用于衡量模型在实验数据上的预测程度或拟合程度。虽然拟合优度指标没有考虑到由于模型的复杂度增大而带来的过拟合的影响，但它在认知建模中的作用也不可忽视。首先，拟合优度指标可以用于探究模型的绝对性能，其次，拟合优度的指标可以在模型的复杂度相差不大以及存在嵌套模型的情况下被用于比较各个模型。在认知建模领域里常用的拟合优度指标包括如下：平方误差(Mean squared error)、决定系数(Coefficient of determination, $r^2/pseudo\ r^2$)、对数似然函数(Log likelihood function)、接收者操作特征曲线(Receiver operator characteristic, ROC)和后验预测检查(Posterior predictive check)。需要注意的是，MSE、r^2不适用于比较本示例数据中的两个模型，因此将仅做文字介绍。

3.1 平均平方误差

平均平方误差，简称为 MSE(Mean squared error)，又称均方偏差(Mean squared deviation, MSD)，是评估一般线性回归的常用指标，其计算公式为：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

其中，y_i是样本的数据点，$\hat{y}_i$是模型的预测值。MSE 通常应用于建模数据是连续变量的回归预测问题中。MSE 并不适用于如本文案例一样的分类问题。

对 MSE 开根号可得到均方根误差(Root mean square deviation, RMSD)；给 MSE 乘以数据点数量，可得到残差平方和(Residual sum of squares, RSS)。当模型使用高斯分布时，RSS 可用于嵌套模型的 F 检验。嵌套模型指的是存在一个完整模型和一个简单模型。简单模型是完整模型的特例，相比于完整模型，简单模型缺少某个参数或者该参数被固定到一个值。

F 值公式为：

$$F = \frac{\frac{RSS_{Reduced} - RSS_{Full}}{\Delta p}}{\frac{RSS_{Full}}{dF_{Full}}} \tag{3}$$

上式中$RSS_{Reduced}$和RSS_{Full}分别为简单模型和完整模型的 RSS，Δp为二者的自由参数之差，dF_{Full}为完整模型的自由度(Hair et al., 2010)。除此之外，高斯分布的 RSS 还可以在计算 AIC 和 BIC 时替代对数似然函数(Friedman et al., 2001; Lebreton et al., 2019)。更多关于 AIC 和 BIC 的内容请分别参考下文 4.1 和 5.1 节。

3.2 决定系数

决定系数 r^2常被用于衡量线性回归模型的拟合优度，r^2的值介于 0 到 1 之间，反映了因变量的变异能被自变量所解释的占比。r^2越接近于 1，模型对数据的拟合效果越好。其计算公式为：

$$r^2 = 1 - \frac{RSS}{TSS} \tag{4}$$

TSS(Total sum of squares)为总平方和，RSS(Residual sum of squares)与为残差平方和，他们的计算公式为：

$$TSS = \sum(y_i - \bar{y})^2 \tag{5}$$

$$RSS = \sum(y_i - \hat{y}_i)^2 \tag{6}$$

与 MSE 一样，决定系数r^2常应用于建模变量为连续变量的回归预测问题，并不适用于本文案例中建模数据为离散分布的分类问题。

为了让r^2也适用于离散分布的情况，研究者提出使用$pseudo\ r^2$。$pseudo\ r^2$有多种计算公式，本文以 McFadden (1984)提出的一种为例进行介绍，因为它较为符合 Kvålseth (1985)提出的八种决定系数应有的性质(Menard, 2000)。

其公式：

$$pseudo\ r_{McFadden}^2 = 1 - \frac{\sum LLF_{Full\ model}}{\sum LL\ F_{Null\ model}} \tag{7}$$

$\sum LLF_{Full\ model}$为模型的对数似然函数之和；$\sum LLF_{Null\ model}$为空模型(空模型指的是参数为 1/选项数量的多项式模型)的对数似然函数之和(Daw, 2011; McFadden, 1984)。在示例数据中，模型一的$pseudo\ r^2$为 0.814，模型二的$pseudo\ r^2$则是 0.803。这说明这两个模型对数据的绝对拟合程度均良好，但模型一比模型二更好。

3.3 对数似然函数

对数似然函数是给定了参数的情况下，模型预测当前数据的概率，反映模型与实际数据的匹配程度。通常在极大似然法估计(Maximal likelihood estimation, MLE)里使用，其公式为：

$$\log \mathrm{L}(\theta|\mathrm{y}) = p(y|\theta) \tag{8}$$

不同任务的对数似然函数不尽相同。当建模数据是选项数据时，对数似然函数通常是伯努利分布或者多项式分布；而建模数据是反应时或者肌电等，对数似然函数则一般为高斯分布(Ballard et al., 2019; Ikink et al., 2019; Li et al., 2011)。

在认知建模的模型比较中，对数似然函数通常有两种用途。第一，使用平均对数似然函数来探究模型绝对的表现(Absolute performance)。本文的示例为二选项任务(Binary choice task)，个体随机选择的概率为 50%，其对数为-0.693。因此当平均对数似然函数大于-0.693 时，模型的表现要优于随机水平(Chance level)。

第二，对数似然可用于计算似然比检验(Likelihood-ratio test)，来推断嵌套模型之间的表现差异是否显著。似然比检验的渐近分布为卡方分布，其自由度正比于两个模型中自由参数数量之差(Casella & Berger, 2002; Wilks, 1938)。

似然比检验的公式为：

$$LRT = -2 \times (\log L_{Reduced} - \log L_{Full}) \tag{9}$$

其中L_{Full}是完整模型的似然函数，$L_{Reduced}$则是固定某些参数的模型的似然函数。具体计算时，我们需要将所有被试的全部试次的似然函数相加，以此计算 LRT，并通过检查卡方分布判断模型差异是否显著。在本文的案例中，模型一和模型二的自由参数数量之差为 2，再乘上被试数量 10，因此，可用自由度为 20 的卡方分布来进行似然比检验。模型一和模型二的似然比检验的 p 值为$1.81e^{-33} < 0.001$，说明二者的拟合差异显著。

3.4 ROC 曲线

ROC 曲线是一种用于评估二分类模型的方法，在信号检测论有着广泛的应用。ROC 曲

线根据不同的分类阈值进行绘制，反映了在不同反应阈值下击中率(True Positive Rate，TPR)与假阳性率(False Positive Rate，FPR)之间的关系(Bishop, 2006)。在 ROC 曲线里，其横坐标为假阳性率，纵坐标为击中率。

在 ROC 曲线里，TPR 是指正确分类的正例数与所有实际正例数之比。FPR 则是指错误分类为正例的负例数与所有实际负例数之比。这里的正例即正确的反应，也即信号检测论的信号，而负例则为错误反应，信号检测论中的噪音。为了绘制 ROC 曲线，我们需要变化反应阈值，计算不同反应阈值下的假阳率和击中率。

ROC 曲线展示了在不同反应阈值下模型的性能。而 AUC(Aera under curve)则衡量了 ROC 曲线下的面积。AUC 的值介于 0 和 1 之间，表示分类器在区分正例和负例方面的能力。AUC 为 0.5 时模型的预测是随机的。而 AUC 的值越接近 1，表示分类器性能越好。一般情况下，当 AUC 大于 0.8 时，我们可以认为模型的性能表现较佳。在示例数据中，模型一的 AUC 面积为 0.956，模型 2 的 AUC 面积为 0.951，二者的 AUC 面积均较大（见图 4）。

ROC 曲线在正负样本大小均衡时表现良好，但是当正负样本差异较大时，ROC 的结果误差极大。当样本不均衡时，查准率-查全率曲线(Precision-recall curve, PRC)是更适合的指标(Davis & Goadrich, 2006)。并且，ROC 曲线仅限于二分类问题，在多分类问题时，绘制 ROC 曲线需要把多分类问题简化为二分类问题(一对多比较或者遍历所有的两两比较等等)(Allwein et al., 2001)。

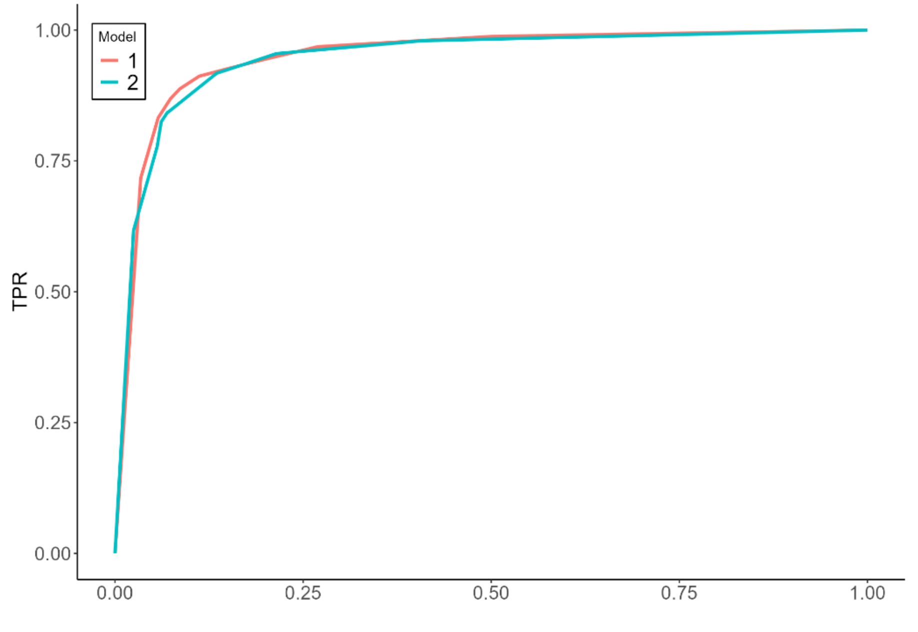


图 4. 案例中模型 1 和模型 2 的 ROC 曲线

3.5 后验预测检查

后验预测检查（posterior predictive check）通常并不属于模型拟合优度，但考虑到该方法也可以用于衡量模型对于原始数据的拟合程度，因此本文将其视为模型拟合优度的指标中的一类。

后验预测检验属于模型验证的方法(Model validation)，检查了模型对样本数据的重现能力(Palminteri et al., 2017; Steingroever et al., 2014; Vandekerckhove et al., 2011)。其公式为：

$$p(y_{rep}|y, M) = \int p(y_{rep}|\theta, M)p(\theta|y, M)d\theta \quad (10)$$

其中M是模型，y是样本数据，y_{rep}是模型重现的样本数据(Gelman, Carlin, et al., 2013; Zhang et al., 2020)。

在实际应用中，后验预测检查的流程如下：在拟合完模型并得到拟合参数后，将拟合的参数代入到模型之中，生成模拟数据。然后通过绘图或者计算一些统计指标(如 MSE 等)来比较模型模拟的数据和真实数据的差异，以评估模型的拟合效果和预测能力(van de Schoot et al., 2021)。

后验预测检查能避免只使用模型比较指标时可能的问题。例如，Palminteri et al. (2017)通过一个模拟研究证明，假设有两个模型 A 与 B，即使在多数情况下，模型 A 的模型选择的指标均优于模型 B，但是模型 A 却有可能无法模拟出数据的总体变化趋势，而 B 模型却可以。因此，除过传统常见的拟合优度指标之外，模拟数据对评估模型来说是至关重要的。

尽管后验预测检查是贝叶斯统计中的概念，但并不代表它仅适用于贝叶斯参数估计。对于非贝叶斯参数估计的模型，我们只能获得参数的点估计，但是我们仍可以使用点估计的参数模拟数据，再将其与真实数据进行对比。虽然在过去的计算模型研究中，后验预测检查并没有得到广泛应用，但在当今越来越多的研究中，研究人员选择使用后验预测检查来评估模型。可以预见，在未来的研究中，后验预测检查有可能成为必不可少的步骤之一(Zhang et al., 2020)。

4 交叉验证类的指标

交叉验证是机器学习领域中用于检验模型对于样本外数据的泛化能力的基本方法。然而，在心理学领域中，直到最近才开始重视这一方法(Daniel et al., 2020; Verstynen & Kording, 2023)。交叉验证的流程包括，首先将数据集分为训练集(Training set)和验证集(Validation set)；然后在训练集上拟合不同的模型；最后在验证集上对比不同模型的预测准确度，从而选择出最优模型(Friedman et al., 2001; Geisser & Eddy, 1979)。

交叉验证主要有三个优点。第一，与许多建立在假设和推导上的指标相比，交叉验证利用计算机的算力替代复杂的推导，使得它极为简洁和直观。第二，交叉验证在权衡模型拟合优度和复杂度时自然地将三种模型复杂度因素(参数数量、参数空间范围和数学形式)考虑在内，而这是许多指标所不具备的。第三，交叉验证不仅可以作为模型选择的相对指标，还可结合前文提到的 MSE、AUC 等统计指标，评估模型数据分布的拟合能力。

常见的交叉验证方法包括 K 折交叉验证(K-fold cross-validation)和留一法交叉验证(Leave-one-out cross-validation)等。K 折交叉验证把数据分成 K 分，其中 K-1 份数据作为训练集，剩余一份数据作验证集。留一法交叉验证则是 K 折交叉验证的特例，它从数据集中每次取出一个样本作为测试集，剩余样本作为训练集。例如，在 N 个样本点的数据集，N-1 个数据样本将作为训练集，而剩下的一个样本是验证集，即$K = n$。留一法交叉验证需要进行 N 次评估才能完成对所有数据样本的预测，因此它的计算量较大。当样本数据噪音较少的情况下，留一法能做到至少与任意 K 值的 K 折交叉验证相同的表现；而当样本数据噪音较多的情况下，留一法的泛化误差则较大(Zhang & Yang, 2015)。

尽管交叉验证是机器学习领域最为常用的验证模型泛化能力的手段，但是交叉验证在认知建模领域里的使用并不广泛，主要原因在于留一法交叉验证的计算量往往较大，而 K 折交叉验证则面临着把数据分为几份的问题。考虑到数据样本量的限制以及计算复杂性，认知建模的研究者往往使用信息准则的近似的指标去代替交叉验证的指标。本文在这里介绍四类常见的指标，分别为 AIC、DIC、WAIC 和 PSIS-Loo-CV。

4.1 AIC

AIC(Akaike information criterion)是最早的模型比较指标之一(Akaike, 1974)，有着详实的理论基础。首先，AIC 是模型所预测的数据分布与真实数据分布的差异。其次，AIC 还被证明是对样本外预测能力(Out-of-sample predictive accuracy)和 LOO-CV 的近似(Stone, 1977)。

AIC 的计算公式为：

$$\mathrm{AIC} = -2 \times \log L(\hat{\theta}|y) + 2 \times K \tag{11}$$

其中，$\log L(\hat{\theta}|y)$是使用极大似然法估计或者最大化后验概率估计求得的最优参数$\hat{\theta}$的对数似然函数值, 可以参考 0 节；K为参数数量，用于对模型复杂度的惩罚。AIC 的值越小，表明模型的拟合效果越好。

因为 AIC 在较小的样本数据中可能会表现不佳(Sugiura, 1978)，有研究者提出基于小样

本偏差修正的 AICc(Hurvich & Tsai, 1989)。AICc 的计算公式为：

$$AIC_C = -2 \times \log L(\hat{\theta}|y) + 2 \times K \times \left(\frac{n}{n-K-1}\right) = AIC + \frac{2\times K\times(K+1)}{n-K-1} \tag{12}$$

其中n是样本数量。AICc 在样本量较大时会趋近 AIC。当样本量较小时，AICc 对复杂的模型的惩罚大于 AIC。Anderson and Burnham (2004)建议当n/K小于 40 时使用 AICc，而当n/K大于 40 时，使用 AIC 和 AICc 则无太大差异。在认知建模领域，由于行为实验中被试完成的试次数量有限，AICc 往往是比 AIC 更合适的指标(Li et al., 2020; Li & Ma, 2021; Suzuki et al., 2012)。

对于 AIC 的差异在多大时才能证明一个模型优于他者的问题，Burnham and Anderson (2004)的建议是，当两个模型的 AIC 之差绝对值小于 2 时，两个模型之间几乎无差异；该值在 4 到 7 之间时，存在较少的证据支持 AIC 值更小的模型；该值大于 10 时，则有充足的证据认为 AIC 小的模型是最优模型。此外，AIC 渐进于卡方分布(Anderson & Burnham, 2004)，因此，研究者可以使用卡方检验对比不同模型的 AIC 值是否存在显著差异。

AIC 的另一个作用在于它可以转换成模型概率，得到所谓的赤池权重(Akaike weight)(Wagenmakers & Farrell, 2004)。

假设有 N 个模型，1 第 i 个模型的赤池权重计算公式如下：

$$\Delta AIC_{M_i} = AIC_{M_i} - minAIC \tag{13}$$

$$w_{M_i} = \frac{\exp(-0.5 \times \Delta AIC_{M_i})}{\sum_{n=1}^{N} \exp(-0.5 \times \Delta AIC_{M_n})} \tag{14}$$

Anderson and Burnham (2004) 认为赤池权重是对下文介绍的后验模型概率 (Posterior model probability, PMP) $p(M_i|y)$的近似，代表在给定样本数据的情况下，模型被选择成为候选模型中最优模型的概率。

AIC 在认知建模中的应用格外广泛，但是它也具有一些缺陷。第一，作为对样本外预测能力的近似，AIC 的精确度不如后续将介绍的 WAIC 和 PSIS-Loo-CV 等指标。其次，AIC 在推导过程中使用插入预测(Plug in prediction)概率$p(y_{rep}|\hat{\theta})$评估模型在样本内的预测准确度，而不是对完整的预测分布进行评估，导致对样本外数据的预测有一定的偏差。最后，AIC 衡量模型复杂度时只考虑了参数数量，忽略了 Myung and Pitt (1997)总结的影响模型复杂度的另两个因素。

4.2 DIC

DIC(Deviance information criterion)是最常见的贝叶斯统计的模型选择指标之一，其理论基于贝叶斯模型样本外预测能力(Expected log pointwise predictive density for a new dataset,

elpd)，DIC 是对 elpd 的近似，因此 DIC 也只适用于贝叶斯参数估计的模型。

DIC 通常被认为是贝叶斯参数估计版的 AIC，但是与 AIC 不同的是 DIC 仅适用于基于 MCMC(Markov chain Monte Carlo)采样估计的模型(Spiegelhalter et al., 2002)。

DIC 的计算公式为$DIC = D(\bar{\theta}) + 2 \times p_D$。DIC 用模型分布与真实模型分布的偏差(Deviance)来衡量模型的性能。偏差的公式为：

$$D(\theta) = -2 \times \log\ L(y|\theta) \tag{15}$$

DIC 的公式的第一项是偏差的后验均值，是模型拟合的好坏代表，其计算公式为：

$$\bar{D}(\theta) = -2 \times \left(\frac{1}{S}\sum_{s=1}^{S} \log L(y|\theta_s)\right) \tag{16}$$

其中S是 MCMC 的采样数。DIC 公式的第二项p_D被称作有效参数，起到了对更为复杂的模型的惩罚作用。其计算公式为：

$$p_D = \bar{D}(\theta) - D(\bar{\theta}) \tag{17}$$

$$D(\bar{\theta}) = -2 \times \log L(y|\bar{\theta}) \tag{18}$$

除上述公式外，Gelman, Carlin, et al. (2013)也提出了用偏差的方差当作有效参数的方法，其公式为：

$$p_D = 0.5 \times Var(\log\ L(y|\theta)) \tag{19}$$

与 AIC 一样，DIC 值越小的模型拟合的越好。当我们把 DIC 除以-2，即可得到 DIC 对 elpd 的近似。与 AIC 公式中的$2K$(K为参数数量)类似的是，DIC 中的p_D也起到了对更为复杂的模型的惩罚作用。不同的是，DIC 里的p_D不仅考虑了模型参数数量，同时还对 Myung and Pitt (1997)总结的其他模型复杂度的因素很敏感。因为 DIC 的这一特性，它时常能带给研究者更多的理解。例如，LBA(Linear ballistic accumulator)模型与 DDM(Drift-diffussion model)同属于对反应时建模的序列抽样模型(Brown & Heathcote, 2008)。LBA 通常被认为是 DDM 的简化版，为验证这二者谁更复杂，Donkin 等人使用 DIC 对二者进行了对比(Donkin et al., 2009)。结果发现，尽管 LBA 模型的参数数量比漂移扩散模型更少，但是 LBA 模型 DIC 指标中p_D更大，这表明 LBA 模型可能并没有简化 DDM。

首先，贝叶斯参数估计的先验为有信息且合适的先验时，能降低模型过拟合的程度。相较于频率主义统计，贝叶斯参数估计更适合构建分层模型，可以同时对所有被试的数据进行拟合，使得模型拟合的结果更少出现极端值(Ahn et al., 2017; Gelman, Carlin, et al., 2013)。其次，DIC 对样本外预测能力的近似比 AIC 更精确。最后，相较于 PSIS-Loo-CV 而言，DIC 的计算简便，常用的 MCMC 软件如 Winbugs(Ntzoufras, 2011)和 Jags(Plummer et al., 2016)均

内置了 DIC 的计算方法(Myung & Pitt, 2018)。

DIC 同时也有不少的缺点。例如 DIC 的表现受参数后验分布的形态以及参数点估计的稳定性的影响较大。当参数后验分布的点估计不能很好地用均值代表，或者模型参数为非指数族分布时，DIC 的估计可能存在偏差。例如，当参数后验分布呈多峰时 DIC 均容易小于 0(Evans et al., 2020; Spiegelhalter et al., 2014)。

4.3 WAIC 和 PSIS-Loo-CV

WAIC(Widely applicable information criterion)(Watanabe, 2010)和 PSIS-Loo-CV(Pareto smoothed importance sampling-leave-one-out cross-validation)(Vehtari et al., 2017)与前面介绍的 DIC 类似，是对 elpd 的近似，且也仅适用于基于 MCMC 采样的贝叶斯模型。

与 DIC 不同，WAIC 使用了 lpd(Log pointwise predictive density, 也在一些文章中缩写为 lppd)去近似 elpd。lpd 是模型在当前样本数据点上模型的预测力，其计算公式为：

$$\widehat{lpd} = \sum_{i=1}^{n} \log\left[\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right] \tag{20}$$

其中，i 为各个数据点，S 为 MCMC 采样的数量。通过 lpd 近似 elpd 时往往会高估 elpd，即高估模型的预测能力。因此，WAIC 在计算 elpd 时引入了一修正项$\hat{p}_{waic}$，这一项与 AIC 里的参数数量和 DIC 里的p_D类似，都是用于惩罚模型的复杂度。$\hat{p}_{waic}$代表估计出的参数的有效数量(estimated effective number of parameters)，其计算公式为：

$$\hat{p}_{waic} = \sum_{i=1}^{n} Var_{s=1}^{S}(\log \mathrm{p}(y_i|\theta^s)) \tag{21}$$

$$\widehat{elpd_{waic}} = \widehat{lpd} - \hat{p}_{waic} \tag{22}$$

为了使 WAIC 渐进于卡方分布，我们可以将其乘上-2。值得注意的是，$\widehat{elpd_{waic}}$越大，模型的样本外预测能力越好，而 WAIC 越小说明模型拟合越好。

与 DIC 相比，虽然 WAIC 也采用插入预测的方法来评估样本外泛化能力，但是 WAIC 具有额外的多个优势。第一，WAIC 利用整个后验分布计算模型复杂度的惩罚项，其结果更稳定。第二，WAIC 在参数后验分布为非正态的模型上的表现也要优于 DIC(Myung & Pitt, 2018)。

贝叶斯留一法交叉验证(Bayesian leave-one-out cross-validation)也可以被用于近似 elpd。

其计算公式为：

$$elpd_{loo} = \sum_{i=1}^{n} \log p(y_i|y_{-i}) \tag{23}$$

$$p(y_i|y_{-i}) = \int p\,(y_i|\theta) \times p(\theta|y_{i-1})d\theta \qquad (24)$$

其中 i 是数据样本点。基于 $elpd_{loo}$ 的信息准则指标为 Looic(Leave-One-Out Cross-Validation Information Criterion)，是$elpd_{loo}$乘以-2。对于留一法交叉验证来说，其对模型复杂度的惩罚项为$elpd_{loo}$和$\widehat{lpd}$之间的差异。

贝叶斯留一法交叉验证计算量极大。为了简便计算，Vehtari et al. (2017)提出了 PSIS-Loo-CV 去近似完整的 Loo-CV。PSIS-Loo-CV 使用了 MCMC 样本，大幅度降低了计算量。因为 R 语言中 loo 包纳入了该算法，这使得它被广泛应用于实际研究中。此外，PSIS-Loo-CV 提供了一项模型诊断指标：帕累托分布的参数 k 值，若绝大多数数据点的 k 值大于 0.7，则说明模型的设置可能存在问题。

除了使用 WAIC 和 PSIS-Loo-CV 进行模型比较外，Vehtari et al. (2019)还推荐使用结合 PSIS-Loo-CV 和集成学习里的堆叠(Stacking)方法(Friedman et al., 2001)去计算每个模型的权重，具体细节可见 Yao et al. (2018)。与赤池权重一样，堆叠方法的模型的权重可用于模型平均。值得注意的一点是，当堆叠方法的模型权重用于模型比较时，表现相似的两个模型会互相“分享”权重，导致二者权重较低且相近(Sivula et al., 2020)。

与 WAIC 比起来，PSIS-Loo-CV 被证明是对 elpd 更好的近似(Vehtari et al., 2016)，使得 PSIS-Loo-CV 能更全面地考虑 Myung and Pitt (1997)提出的三个影响模型复杂度的因素。并且 Vehtari et al. (2017)开发的 R 包 loo 降低了使用门槛，研究者只需要输入 MCMC 采样的似然函数，即可计算 WAIC 和 PSIS-Loo-CV。关于使用 WAIC 和 PSIS-Loo-CV 的具体建议，可以详见 Vehtari (2022)。

4.4 不同交叉验证近似指标的总结

交叉验证类的指标在认知建模中的使用极广，随着近年来黑箱 MCMC 软件的流行，使得研究者能较为容易地使用贝叶斯参数估计，这极大地推广了 DIC、WAIC 和 Loo-CV 的使用。因为交叉验证类的指标更容易确认复杂模型的为最优模型，这使得它们在心理学研究的应用格外的广泛。

虽然上述这些指标建立在不同的假设和近似方法的基础之上，AIC 更多地应用在极大似然法估计或者最大后验概率法拟合的模型，而 DIC、WAIC 和 Loo-CV 则用于 MCMC 估计的模型中。但是在一些认知建模的应用里，它们的差异并不明显。例如，Evans (2019) 在 LBA 模型上对比了 AIC、DIC、和 WAIC，虽然它们的表现类似，但是 DIC 和 WAIC 的表现要略优于 AIC。又比如，Westbrook et al. (2020)使用和 AIC 和 DIC 对比了不同的注意力

DDM(Attentional drift-diffusion model, aDDM)，二者的结果几乎一致。

在本文的案例里，我们用最大化后验概率法的结果计算了 AIC，并用分层贝叶斯参数估计的结果计算了 DIC、WAIC 和 PSIS-Loo-CV，如图 5。对于贝叶斯模型比较指标，根据 Vehtari et al. (2017)，我们可以对不同模型进行 Wald 检验，从而判断模型之间是否有显著的差异。Wald 检验的结果表明，两模型的 DIC 存在显著差异，$D_{DIC} = 25.03 > 1.96 \times \sigma_{DIC} = 22.85$，其中 D 表示模型 2 与模型 1 在交叉验证指标上的差异。而 WAIC 和 PSIS-Loo-CV 的表现几乎一致，模型之间的差异也显著，$D_{WAIC/Loo-CV} = 22.70 > 1.96 \times \sigma_{WAIC/Loo-CV} = 21.56$。

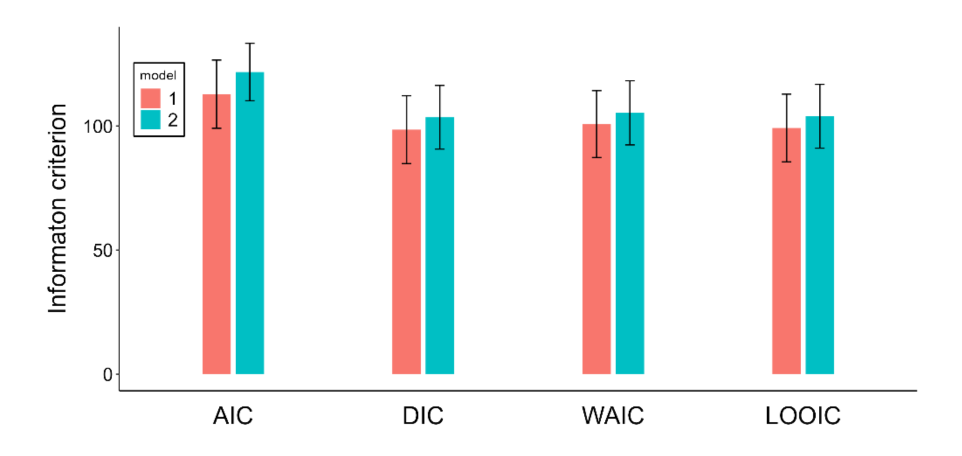


图 5. 不同交叉验证类的近似指标对模型一和模型二的评估，信息准则指标越小代表模型拟合的越好。

注：PSIS-Loo-CV 计算的结果常记作 LOOIC(Leave-One-Out Information Criterion)。

5 边际似然

边际似然或称作模型证据则是另一大类的模型评估指标，同时也是贝叶斯模型选择(Bayesian model selection, BMS)的核心。贝叶斯参数估计的公式为：

$$p(\theta|y) = \frac{p(y|\theta)\times p(\theta)}{\int p(y|\theta)\times p(\theta)d\theta} \quad (25)$$

但是上式忽略了模型 M 这一项。如果对上式进行修改，增加 M，即可得：

$$p(\theta|y,M) = \frac{p(y|\theta,M)\times p(\theta,M)}{\int p(y|\theta,M)\times p(\theta,M)d\theta} \quad (26)$$

此时贝叶斯公式中的分母即为模型的边际似然或模型证据。边际似然计算的是参数空间范围内模型对数据的平均拟合(Average fit)，边际似然越大，模型对样本数据解释的越好。

边际似然可以平衡模型的复杂度和拟合效果。例如，较简单的模型可能具有较低的拟

合优度，但是却有较高的边际似然，因为它们的参数空间不确定性小。相反，复杂的模型可能具有较高的拟合优度，但是其边际似然却较小，因为它们在参数空间的不确定性较大(MacKay, 2003)。

边际似然同时考量了 Myung and Pitt (1997)总结的三种影响模型复杂度的因素，如图 6 所示。过于简单的模型给予观测数据的概率$p(M|y)$往往很少，因此其边际似然也很小；过于复杂的模型的数据分布更广，但是它分给当前观测数据的概率$p(M|y)$也很小，由此其边际似然也较小；只有当复杂度适中时，观测数据对应的边际似然才会较大。

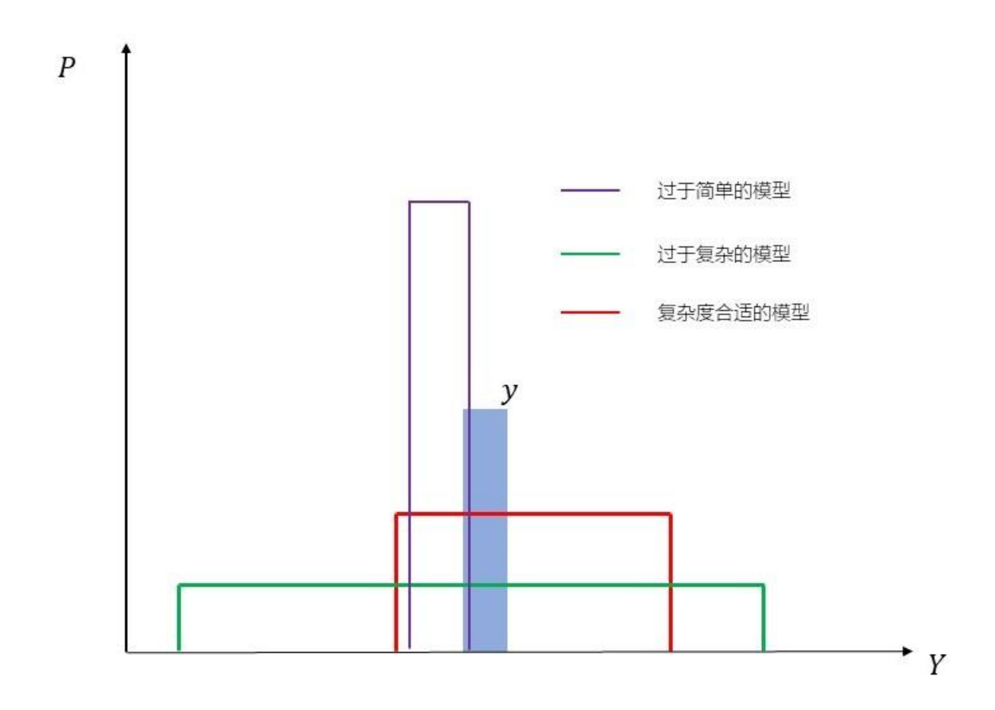


图 6. 边际似然对不同类模型的惩罚。横坐标为数据值；纵坐标代表数据值对应的似然值。

边际似然还对贝叶斯参数拟合的先验信息格外地敏感。例如，当使用弱信息的先验分布时，复杂模型的边际似然小于简单模型；当使用更窄的、信息更丰富的先验分布时，复杂模型的边际似然就有可能大于简单模型(Farrell & Lewandowsky, 2018)。

边际似然在实际的应用中存在两个主要问题。第一，先验分布对边际似然的计算结果有较大的影响。当我们的数据点较多时，先验分布对参数估计的结果不恰当的先验分布会对边际似然的计算结果产生很大的影响(Boehm et al., 2018)。对于先验的选择，主观贝叶斯方法认为应当根据已有的知识和信念选择先验分布，而客观贝叶斯方法则试图排除先验选择的个人因素，更多地使用如先验杰佛里斯默认先验分布(Jeffreys default prior distribution)等无信息的先验分布(Jeffreys, 1998; Vandekerckhove et al., 2015)。为了选择出更合适的先验分布，研究者可以使用敏感性分析(Prior sensitivity check)，变换不同的先验分布检查其对边

际似然的影响。

第二个问题是，计算边际似然需要对先验分布和模型的似然函数在整个参数空间上进行乘积积分。然而只有极少的简单模型的边际似然可以直接求解，更多模型的边际似然是无法简单计算的。因此，许多近似方法和采样积分方法被提出以用于计算边际似然。

5.1 BIC

BIC(Bayesian information criterion)(Schwarz, 1978)与 AIC 类似，也是为最经典、应用最为广泛的模型选择指标之一。BIC 是下文中拉普拉斯近似(Laplace approximation)边际似然的一个特例(Bishop, 2006)。当计算拉普拉斯近似时，假设先验分布为无信息先验，且当数据点 n 的数量极多时，根据大数定律，拉普拉斯近似计算的结果可以被简化为 BIC。

BIC 的计算公式为：

$$BIC = -2 \times \log L(\hat{\theta}|y) + K \times ln(n) \tag{27}$$

其中，$Kln(n)$是 BIC 里对模型复杂度的惩罚项，K是参数数量，n是数据的数量。可见，BIC 不仅考虑了模型参数数量对惩罚模型复杂度的影响，也将数据量作为惩罚模型复杂度的关键因素，BIC 与 AIC 一样，其值越小说明模型拟合的越好。

除此之外，BIC 有根据样本矫正的 SABIC(Sample-adjusted BIC)(Sclove, 1987)，然而其缺乏理论依据，应用较少(Dziak et al., 2020)。

虽然 BIC 是最常见的模型选择指标(Wilson & Collins, 2019)，然而 BIC 仍然存在缺点。第一，BIC 对模型复杂度的惩罚只考虑了模型的参数和样本数量，并没有考虑到 Myung and Pitt (1997)总的另外两个影响模型复杂的因素，即参数空间范围和模型的数学形式。第二，虽然 BIC 是在贝叶斯理论的框架下推导而来，但是它并未考虑不同先验信息对结果的影响。

5.2 近似方法计算边际似然

本文介绍的近似方法计算边际包括 Savage-Dickey 比(Savage-Dickey Ratio, SDR)、拉普拉斯近似(Laplace approximation)，核密度估计方法(Kernel density estimation, KDE)以及变分推断。与 BIC 相比，这些方法考虑了先验分布的影响，但其计算量并没有显著增大；与后文介绍的采样方法相比，近似方法的误差更大，但其计算量却远小于采样方法，使得它在很多研究中得到了应用。

Savage-Dickey 比适用于在嵌套模型的模型比较中计算二者的贝叶斯因子(Dickey, 1973; Dickey, 1976; Wagenmakers et al., 2010)。假定简单模型所缺少的参数为θ，Savage-Dickey 比将嵌套模型的贝叶斯因子计算简化为完整模型θ等于 0 时的后验概率与先验概率之比。在本

文的案例中，当我们将负责巴普洛夫效应的两个参数 π 和 b 的组水平的均值参数固定为 0 时，计算对数贝叶斯因子为 2.24。。Savage-Dickey 比适用于各个参数的先验分布是相互独立的情况，但当先验分布是有协方差矩阵的多维分布时则需要矫正(Heck, 2019)。

拉普拉斯近似主要应用于使用最大化后验概率拟合模型的情况，其主旨在于使用多维高斯分布来近似参数的分布，并用泰勒展开避免积分问题。与 BIC 相比，拉普拉斯近似的边际似然考虑了先验分布的影响，且其计算误差更小。拉普拉斯近似的计算边际似然的公式为：

$$\log p(y|M) \approx \log L(\hat{\theta}|y) + \log p(\hat{\theta}|M) + \frac{K}{2} \times \log 2\pi - \frac{1}{2}\log|H| \quad (28)$$

其中$|H|$为负对数后验的海森矩阵行列式。拉普拉斯近似是心理学里最常见的近似计算边际似然的方法之一(Gershman, 2016; Huys et al., 2011; Myung & Pitt, 1997)，其关键步骤在于计算海森矩阵的行列式，但当海森矩阵为非正定矩阵时，$\log|H|$这一项有可能为非数值(NaN)。

核密度估计方法则可利用 MCMC 采样得到的参数后验分布来计算边际似然。核密度估计方法使用了非参统计方法中的核密度估计计算参数的后验概率$p(\hat{\theta}|y) = k(\hat{\theta}|\theta, \phi)$。其中，$k$为密度核函数，通常为高斯分布(Wasserman, 2006)。θ是 MCMC 采样获得的各个参数样本，而$\hat{\theta}$是 MCMC 采样分布的点估计代表，一般是概率密度最高的点。

在得到了参数的后验概率$p(\hat{\theta}|y)$后，根据贝叶斯公式，我们便可以直接得到边际似然：

$$p(y|M) = \frac{L(\hat{\theta}|y) \times p(\theta|M)}{p(\hat{\theta}|y)} \quad (29)$$

核密度估计方法计算简便，且不受海森矩阵的限制，一些模拟研究还发现它的表现要比拉普拉斯近似等方法更好(Bos, 2002)。

变分推断(Varitional inference)是除采样方法外另一常见的贝叶斯参数估计的方法。与采样方法不同的是，变分推断试图用变分分布$q(z)$近似参数后验分布$p(\theta|D)$，从而将贝叶斯公式里的积分问题变换成优化问题(Bishop, 2006)。变分推断不仅仅在贝叶斯参数估计里有着许多应用，它还可以被当作理解认知过程的理论(Friston et al., 2006)。

变分推断的优化函数被称作证据下界 ELBO(Evidence Lower Bound)或者负自由能(Negative free energy)(Bishop, 2006; Friston et al., 2007)，是对数边际似然的下限。最大化 ELBO 时能获得边际似然的估计值，ELBO 的公式为：

$$ELBO = E_{q(z)}[\log\frac{p(\theta, y|M)}{q(z)}] = E_{q(z)}[\log p(y|\theta, M)] + D_{KL}(q(z)||p(\theta|M)) \quad (30)$$

ELBO 的公式表明边际似然可以被分为两部分，第一部分是似然函数在变分分布上的

期望值，代表模型拟合的好坏；第二部分是变分分布和先验分布的 KL 散度，代表后验和先验的差异。当模型拟合程度越差或者先验分布与后验分布之间的差异越大时，边际似然越小(Stephan et al., 2009)。

在实际应用里，基于 Matlab 的变分推断的工具包 VBA 在拟合模型完毕时可以返回优化 ELBO(Daunizeau et al., 2014)。此外，基于 Stan 拟合的模型也会返回未标准化的后验分布概率和变分分布概率，可以用于计算 ELBO。变分推断方法问题在于它得到的是边际似然的下限，少有理论研究关注 ELBO 对边际似然的近似误差(Blei et al., 2017)。

5.3 采样方法计算的边际似然

蒙特卡洛采样方法是一种常见的统计模拟的方法，当一个积分公式难以直接求解时，我们可以通过不断地数值采样，带入到公式中计算，逐步逼近积分的结果。因为复杂模型的边际似然的积分无法通过解析解求解，这使得许多蒙特卡洛采样算法被应用到计算边际似然中。

采样方法种类繁多，包括热力学积分(Thermodynamic integration)，序列蒙特卡洛采样(Sequential monte carlo sampler, SMC)和粒子 MCMC 的方法。然而，由于缺少易用的软件，这些方法的应用受到了限制(Doucet & Johansen, 2009; Murphy, 2023)。相比之下，重要性采样(Gamerman & Lopes, 2006; Hammersley, 2013)和桥采样(Bridge sampling)(Gronau et al., 2017; Meng & Wong, 1996)，有着易用的软件或其本身计算简便，广泛应用于心理学研究中。

重要性采样属于蒙特卡洛方法的一种，它的关键在于引入重要性采样分布。当从一个分布里采样困难或者它的样本质量不高时，我们就可以退而求其次，从重要性分布里采样(Bishop, 2006)。在计算边际似然时，我们首先引入重要性采样分布$g_{IS}(\theta)$，从而得到：

$$p(y|M) = \int p(y|\theta, M) \times p(\theta|M) d\theta = \\ \int p(y|\theta, M) \times p(\theta|M) \times \frac{g_{IS}(\theta)}{g_{IS}(\theta)} d\theta = \\ \int \frac{p(y|\theta, M) \times p(\theta_i|M)}{g_{IS}(\theta)} \times g_{IS}(\theta) d\theta = \\ \mathbb{E}_{g_{IS}(\theta)}\left(\frac{p(y|\theta, M) \times p(\theta|M)}{g_{IS}(\theta)}\right) \quad (31)$$

因此，边际似然可由下式得到：

$$\hat{p}(y|M) = \frac{1}{N}\sum_{i=1}^{N} \frac{p(y|\theta_i, M) \times p(\theta_i|M)}{g_{IS}(\theta)}, \tilde{\theta}_i \sim g_{IS}(\theta) \quad (32)$$

通过从重要性分布里不断采样，带入到贝叶斯公式里计算，再将不同样本的结果求和即可得到边际似然。在重要性采样分布里，重要性分布的选择对结果影响极大。为了保证

估计结果有较小的方差，$g_{IS}(\theta)$通常是有一个较厚尾部的分布。此外，当使用重要性采样计算边际似然的倒数$\frac{1}{\hat{p}(y|M)}$时，此时的重要性采样也被称作 RIS(Reverse importance sampling)(Gelfand & Dey, 1994)。相对的，RIS 的采样分布更需要一个有着较薄尾部的分布。

利用 MCMC 采样得到参数后验的样本来计算边际似然能显著减低计算量，此时的重要性采样被称为调和平均估计器(Harmonic mean estimator)。调和平均器易于计算，但是计算结果方差较大，使得它鲜见于实际研究。

提高调和平均估计器性能的常见方法有如下几种。第一，使用加权重要性采样(Acerbi et al., 2018)。此法需要 RIS 乘上一个有着较薄尾部的函数$f(\theta)$，且$\int f(\theta)d\theta = 1$，因此$f(\theta)$可以是多维高斯分布。RIS 计算公式为:

$$\frac{1}{\hat{p}(y|M)} = \frac{1}{N}\sum_{i=1}^{N}\frac{f(\theta_i)}{p(y|\theta_i, M) \times p(\theta_i|M)} \tag{33}$$

第二是将 MCMC 样本替换为均匀分布或者高斯分布与 MCMC 样本的混合分布(Steingroever et al., 2016; Vandekerckhove et al., 2015)，该方法因为便于计算，在心理学有着很多应用。

桥采样是对重要性采样的改善和提升，与重要性采样一样，桥采样也利用了 MCMC 的样本。相较于计算更为简单的重要性采样，桥采样避开了选择分布的步骤，其计算结果的方差更小，并且更适合于分层模型。桥采样的特点在于，通过引入一个连接目标分布和提议分布的桥分布(Bridge distribution)，以此减小计算边际似然的方差并提高计算的精度(Meng & Wong, 1996)。桥采样的缺点在于，其计算较为复杂，需要反复迭代直至结果稳定，这增加了计算的时间和资源，具体可见 Gronau et al. (2017)。Gronau 等人开发的 R 包 bridgesampling 简化了计算过程，使用 JAGS 和 Stan 拟合的模型可以使用该包来计算边际似然。

5.4 不同方法计算边际似然的总结

计算边际似然的方法种类繁多，选择何种方法依赖于具体的使用情景。BIC 是最简单的方法，但它的误差也最大。此外，因为 BIC 是无先验信息的边际似然的近似，理论上使用 BIC 会更倾向于选择更简单的模型。Evans (2019)认为，当研究者使用有信息的先验分布拟合的模型时使用 BIC 是不恰当的。计算边际似然的先验分布应与拟合模型的先验保持一致。

当使用最大化后验概率法拟合模型时，拉普拉斯近似是更简便的方法。如果使用

MCMC采样，且模型非分层模型时，重要性采样、拉普拉斯近似或者KDE方法更为合适，因为它们的计算量更小。若模型是分层模型，此时拉普拉斯近似的海森矩阵的行列式不易计算，再加之重要性采样又面临着采样分布选择的困难，这使得桥采样是更为合理的选择。

当研究者比较两个模型时，可以计算两个模型的边际似然的比值，结果即为贝叶斯因子(Bayes factor) (Kass & Raftery, 1995)。贝叶斯因子的特性在于能够为零假设提供证据，因此它在当前的心理学研究里有着许多应用。关于贝叶斯因子在数据分析的使用，以及其分析结果的解读，可见胡传鹏 et al. (2018)。此外，BIC 作为边际似然的近似，也可以被用于计算贝叶斯因子和后验模型概率(Wagenmakers, 2007)。其计算方法为，将两个模型的 BIC 之差乘以-0.5，然后通过指数函数可以将其转化为贝叶斯因子。

值得注意的是，与常见的数据分析不同，认知建模里贝叶斯因子对比的两个模型可以是任意两个模型，只要它们建模的数据相同即可。而 T-test 和 ANOVA 里对比的两个模型则必须是备择假设和零假设。

在本文的案例里，BIC 和拉普拉斯近似的边际似然均基于最大化后验概率法的结果，我们可以利用两者的结果计算每个被试的贝叶斯因子。相比之下桥采样方法适用于分层贝叶斯估计，可以直接计算组层面的边际似然值，进而可以获得组层面的贝叶斯因子(Group bayes factor, GBF)。

图 7 比较了基于 BIC、拉普拉斯近似和桥采样方法计算组层面贝叶斯因子的结果。需要注意的是，为了方便比较，我们通过求和所有被试在个体层面的贝叶斯因子(基于 BIC 和拉普拉斯近似方法)来获得组层面的贝叶斯因子。结果发现，三种方法下的组贝叶斯因子均支持真实模型，即模型一为最优模型。然而，它们的具体数值差异却极大。BIC 版的对数组贝叶斯因子为 12.59，桥采样版的对数组贝叶斯因子为 39.92，而拉普拉斯近似版的对数组贝叶斯因子值为 50.63。数值的差异不仅是因为不同指标近似的精度不同，同时也受到模型拟合方法差异的影响。

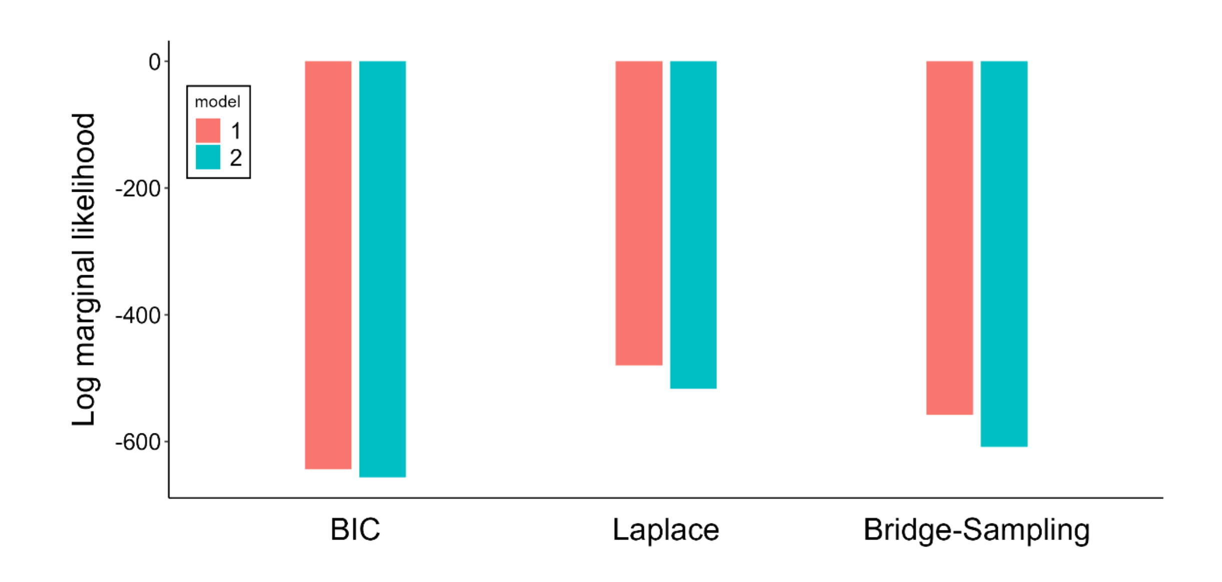


图 7. 不同组边际似然近似指标对模型一和模型二的评估。所有指标均被转换为对数边际似然，其值越大表示模型拟合的越好。

6 总结与展望

计算模型在实验心理学的研究在最近十余年愈发的广泛，而模型比较是认知建模中关键的一环，不恰当地进行模型比较可能会让研究者得出错误的结论。因此，合理地使用模型比较指标对基于计算模型的研究来说至关重要。本文梳理、总结了在认知建模领域常见和新兴的模型选择指标，对最常见的两类指标：基于交叉验证的指标和基于边际似然的指标进行了对比，建议了不同指标的使用条件。并结合一个简单的案例，提供了具体的计算方法。

值得一提的是，过往许多使用计算模型的研究均采用较为简单的模型比较指标，如 AIC 和 BIC 等。这些指标尽管有着许多优点，但却忽视了影响模型复杂度等诸多重要因素。而近年来被推广的指标，诸如 WAIC，近似/采样方法计算的边际似然等较为复杂的指标对模型复杂度的考量要更加的完善，由此基于这些指标的模型比较的结果也更加稳定可靠。随着越来越多成熟且容易操作的工具的发展，这些指标将更多地应用在研究里。

除此之外，早期认知建模的研究大都只注重使用相对指标来评估模型的优劣，忽视了模型拟合的绝对好坏。这导致了一种困境：即便我们选择出了一个最优模型，该模型却并不一定对样本数据有完善的描述。因此，在进行模型比较时，我们首先需要通过相对指标选择最优模型，再通过拟合优度指标评估模型对当前数据拟合的绝对优良度。只有当模型

在相对指标上胜出其他候选模型，且在数据上有着良好的绝对拟合优度时，我们才能将它当作最优模型。随着后验预测检查等方法的普及，今后的研究应将更多地结合相对指标和绝对指标来进行模型评估及模型验证。

6.1 边际似然和交叉验证的争论

本文着重介绍了边际似然与交叉验证这两类最常见的模型比较方法。尽管二者基于的理论大相径庭，但是也有研究表明二者间存在不少联系。例如，Fong and Holmes (2020)证明了边际似然在一些特定情况下与交叉验证等价。但是这二者中哪一个更适合实际研究以及如何选择它们仍有许多在争议。

建模中通常有 M-Closed 和 M-Open 这两种场景。M-Closed 场景假设在候选模型中存在一个“真实”模型，能完美地描述数据的生成过程。M-Open 场景假设所有的候选模型都不能完美地描述数据的生成过程。在 M-Open 场景下，模型选择的目标是找到一个在所有候选模型中表现最好的模型，而不是寻找真实模型(Burnham & Anderson, 2004; Gelman, Hwang, et al., 2013)。

假如在 M-Closed 场景下且数据数量接近无限，此时边际似然能选择出“真实”模型。而在 M-Open 场景下，交叉验证则更适合，它能找出 KL 散度距离“真实”模型最小的模型。虽然在 M-Closed 环境下，交叉验证也能找到与数据 KL 散度最小的模型，但它却无法找出“真实”模型。有研究表明边际似然和交叉验证两者的优势是无法被结合的(Vrieze, 2012; Yang, 2005)。

边际似然的支持者对交叉验证的反驳主要集中在交叉验证无法找出“真实”模型这一点上。例如，Gronau and Wagenmakers (2019)在实验中使用 Beta-Bernoulli 模型生成模拟数据，并使用不同复杂度的模型拟合模拟的数据，最后用 Loo-CV 和基于 Loo-CV 计算的 Pesudo 贝叶斯因子对各个模型进行评估和对比。分析结果发现，除 Loo-CV 会选择复杂度更高的模型而非产生数据的真实模型的固有缺陷外，Loo-CV 对真实模型的支持会随着数据的增长而呈倒 U 型。当数据增长时，Loo-CV 对真实模型的支持会先下降再增长。因此 Gronua 和 Wagenmakers 认为，当研究者使用 Loo-CV 时应该格外谨慎。

Vehtari et al. (2019)驳斥 Gronau and Wagenmakers (2019)的观点，认为 M-Closed 设置只是为了简化建模问题，实际应用中很少出现 M-Closed 环境。并且 Vehtari et al. (2019)认为 Gronau 和 Wagenmakers 错误地使用 Loo-CV 去计算 Pesudo 贝叶斯因子。相反，如果使用堆叠的方法，将各个模型的 Loo-CV 作为输入值，所计算的模型权重可以很好地在 M-Closed

环境下选择出最优模型。

另一方面，交叉验证的支持者们则认为边际似然尽管拥有很多优良的理论特性，但是很多情况其实际应用却不尽如人意。原因在于，边缘似然并不是对模型泛化能力的衡量，而是在给定了先验分布和模型的情况下，衡量模型对当前数据解释的能力。即使一个模型使用了合适的先验分布并具有更好的边际似然，其在样本外数据上的泛化能力也不一定比其他模型更强(Lotfi et al., 2022)。

此外，在贝叶斯推断中，选择合适的先验分布是极为困难的。例如，Gelman, Carlin, et al. (2013)认为，在边际似然的实际应用中，不合适的有信息的先验分布会对边际似然造成极大的影响。模型的先验分布愈是无信息，边际似然的模型比较愈倾向于更简单的模型。而与边际似然相比，Loo-CV 则不会受到这方面的影响(Gelman, Carlin, et al., 2013)。例如，Kennedy et al. (2019)通过对气球模拟风险任务(Balloon Analog Risk Task, BART)实验数据建模，测试了不同先验分布对贝叶斯因子的影响。他们发现，随着先验分布无信息程度的增大，贝叶斯因子会逐渐偏向于简单的模型。在本文的案例中也是如此，因为 BIC 假设了无信息的先验分布，而拉普拉斯近似和桥采样均使用了实际拟合模型的先验分布，所以 BIC 的贝叶斯因子也要远小于其他两者。

6.2 模型选择指标的使用建议

首先，当我们进行模型比较时应当注意每个指标所适用的情况。各个模型比较指标仅适用于与建模数据一致的场景。例如，基于反应时和选项数据的 DDM 的 AIC 无法和基于选项数据建模的强化学习模型的 AIC 进行比较(Fontanesi et al., 2019)。

其次，当模型比较的相对指标无法区分不同的模型时，后验预测检测也可以作为选择模型的方法。例如，Steingroever et al. (2014)发现在爱荷华赌博实验里，BIC 等指标很难区分不同模型，而后验预测检查则能很好地选择出最优模型。

例如，AIC 和 BIC 作为最常见的指标，适用于参数估计方法为点估计的极大似然法的模型，然而如何在 AIC、和 BIC 之间进行选择仍有争议。

BIC 的惩罚项惩罚力度更大，导致它们通常会选择简单的模型。因此，研究者可以根据自己研究假设的效应量和统计功效来选择这些指标。例如，BIC 的一类错误和二类错误都会随着样本量的增大而下降。而 AIC 的二类错误会随着样本量下降，但其一类错误并不会。并且 AIC 的二类错误比 BIC 要小(Dziak et al., 2020)。即在同等样本条件下，AIC 能确认样本外预测能力更好的模型为最优模型，但同时也冒着一类错误更大的风险。而 BIC 虽

然有着确认真实模型的能力，但是其二类错误，即选出一个表现较差的模型的概率也更高。

使用模型复现(Model recovery)的方法来决定究竟使用何种指标也是一种选择(Wilson & Collins, 2019)。例如，Collins and Frank (2012)使用更复杂的模型模拟数据，并用复杂模型和简单模型拟合该数据。他们发现，当使用 BIC 作为模型比较指标时，拟合结果会支持简单模型，也就是说，BIC 往往过于惩罚复杂的模型，导致无法复现出模拟数据背后的真实模型，而 AIC 却可以复现出更为复杂的真实模型(Collins & Frank, 2018)。最后，也有不少研究者推荐同时汇报 AIC 和 BIC。如果二者的结果一致，则模型比较结果也更为可靠。如果二者相悖，则可根据不同的原则进行分门别类的讨论(Farrell & Lewandowsky, 2018)。

除此之外，不同参数估计的方法也会限制模型比较方法的使用。对于使用贝叶斯参数估计的模型而言，我们可以利用 MCMC 样本计算边际似然或者 Loo-CV 等更精确的近似指标。而如果使用了点估计的最大化后验概率法，我们也可以使用拉普拉斯近似计算边际似然。在有信息的先验分布时，边际似然表现会优于 WAIC 等对交叉验证的近似。Evans (2019)使用 LBA 模型对比了不同信息程度的先验分布对模型比较的影响，发现当先验分布是无信息的或者弱信息的时，边际似然倾向于过度成罚复杂模型，导致结果偏离最优选择；而当先验分布是中等程度的信息时，边际似然的结果更接近于最优选择，并且要优于 WAIC；而当先验分布是强信息的时，边际似然又会倾向于选择复杂度过高的模型。因此，当我们对模型的先验有足够的认识并设置有信息的先验时，边际似然可能是更好的选择；当使用无信息先验，或设置有信息的先验但并不确定其是否恰当时，对先验不敏感的 WAIC，DIC 和 Loo-CV 是更恰当的指标。

6.3 模型比较的新发展

传统的模型比较通常要选择出一个最优模型，但单一的模型既可能过拟合，也忽视模型的不确定性。研究者提出贝叶斯模型平均的思路，即同时考虑多个模型影响的权重，以增强基于模型所做出推断的鲁棒性(Clyde et al., 2011; Hinne et al., 2020; Merlise & Edward, 2004)。例如，Boehm et al. (2023)在使用模型平均探究速度-准确性权衡对 DDM 参数的影响时发现，使用贝叶斯模型平均能减少模型过拟合对 DDM 参数估计的影响，使得对 DDM 参数分析的结果更加准确。但值得注意的是，贝叶斯模型平均受限于边际似然的计算，在边际似然计算困难的情况下，难以计算后验模型概率。一种可行的方法是使用赤池权重来替代后验模型概率。此外，结合堆叠方法和 PSIS-Loo-CV 的模型权重也可以用于替代后验模型概率(Yao et al., 2018)。

使用模型比较指标的常见方式是比较指标值在所有被试上的和或者平均值的。然而这种做法忽视被试之间的差异，也忽视了极端值对模型比较的产生的可能影响。源于DCM(Dynamic causal modelling)中模型比较的贝叶斯模型选择(Random effect Bayseian model selection, RE-BMS)(Stephan et al., 2009)能有效地减少极端值的影响，在认知建模中也取得广泛的应用。RE-BMS 利用贝叶斯分层模型来考虑被试的差异，使用了多项式分布和狄利克雷分布以避免数据点非对称分布形态的影响。此外，RE-BMS 引入超出概率(Protected exceedence probability, PXP)，代表在当前样本数据下，某一模型的边际似然大于等于其余模型并可以作为生产当前数据的“真实模型”的概率，即$PXP = p(r_{M_{k=i}} \geq r_{M_{k\neq i}} | y)$。PXP 大于 0.95 就可以像传统的假设检验一样认为该模型要显著地优于其余模型(Iglesias et al., 2013)。值得注意的是，Matlab 中的工具包 SPM、VBA 和 R 中的 bmsR 包均可实现 PXP 的计算(Daunizeau et al., 2014)，使其在认知建模得到广泛应用。此外，当我们使用 AIC、BIC 等信息准则指标作为 RE-BMS 的输入时，需将这些指标除以-2 来保证结果的正确。

参考文献：

胡传鹏, 孔祥祯, & 彭凯平. (2018). 贝叶斯因子及其在 JASP 中的实现. *心理科学进展*, *26*(6), 951-965. https://doi.org/10.3724/sp.J.1042.2018.00951

区健新, 吴., 刘金婷, 李红. (2020). 计算精神病学：抑郁症研究和临床应用的新视角. *心理科学进展*, *28*(1), 111-127. https://doi.org/10.3724/sp.J.1042.2020.00111

Acerbi, L., Dokka, K., Angelaki, D. E., & Ma, W. J. (2018). Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLoS Computational Biololgy*, *14*(7), e1006110. https://doi.org/10.1371/journal.pcbi.1006110

Ahn, W. Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry*, *1*, 24-57. https://doi.org/10.1162/CPSY_a_00002

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, *1*(2), 113-141. https://doi.org/10.1162/15324430152733133

Anderson, D., & Burnham, K. (2004). *Model selection and multi-model inference* (Vol. 63). Second. NY: Springer-Verlag.

Ballard, I. C., Wagner, A. D., & McClure, S. M. (2019). Hippocampal pattern separation supports reinforcement learning. *Nature Communications*, *10*(1), 1073. https://doi.org/10.1038/s41467-019-08998-1

Betts, M. J., Richter, A., de Boer, L., Tegelbeckers, J., Perosa, V., Baumann, V., Chowdhury, R., Dolan, R. J., Seidenbecher, C., Schott, B. H., Duzel, E., Guitart-Masip, M., & Krauel, K. (2020). Learning in anticipation of reward and punishment: perspectives across the human lifespan. *Neurobiology of Aging*, *96*, 49-57. https://doi.org/10.1016/j.neurobiolaging.2020.08.011

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859-877. https://doi.org/10.1080/01621459.2017.1285773

Boehm, U., Evans, N. J., Gronau, Q. F., Matzke, D., Wagenmakers, E.-J., & Heathcote, A. J. (2023). Inclusion Bayes factors for mixed hierarchical diffusion decision models. *Psychological Methods*, Advance online publication. https://doi.org/10.1037/met0000582

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., Krypotos, A.-M., Lerche, V., Logan, G. D., Palmeri, T. J., van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J. J., Voss, A., Wiecki, T. V., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46-75. https://doi.org/10.1016/j.jmp.2018.09.004

Bos, C. S. (2002). *A comparison of marginal likelihood computation methods* Compstat: Proceedings in Computational Statistics,

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153-178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, *33*(2), 261-304. https://doi.org/10.1177/0049124104268644

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Cengage Learning.

Cavanagh, J. F., Eisenberg, I., Guitart-Masip, M., Huys, Q., & Frank, M. J. (2013). Frontal theta overrides pavlovian learning biases. *The Journal of Neuroscience*, *33*(19), 8541-8548. https://doi.org/10.1523/JNEUROSCI.5754-12.2013

Clyde, M. A., Ghosh, J., & Littman, M. L. (2011). Bayesian Adaptive Sampling for Variable Selection and Model Averaging. *Journal of Computational and Graphical Statistics*, *20*(1), 80-101. https://doi.org/10.1198/jcgs.2010.09049

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal Of Neuroscience*, *35*(7), 1024-1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, *115*(10), 2502-2507. https://doi.org/10.1073/pnas.1720963115

Daniel, R., Radulescu, A., & Niv, Y. (2020). Intact Reinforcement Learning But Impaired Attentional Control During Multidimensional Probabilistic Learning in Older Adults. *The Journal of Neuroscience*, *40*(5), 1084-1096. https://doi.org/10.1523/JNEUROSCI.0254-19.2019

Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. *PLoS Computational Biology*, *10*(1), e1003441. https://doi.org/10.1371/journal.pcbi.1003441

Davis, J., & Goadrich, M. (2006). *The relationship between Precision-Recall and ROC curves* Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA. https://doi.org/10.1145/1143844.1143874

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In *Decision making, affect, learning: Attention performance XXIII* (Vol. 23).

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, *19*(8), 1153-1160. https://doi.org/10.1016/j.neunet.2006.03.002

Dickey, J. (1973). Scientific Reporting and Personal Probabilities: Student's Hypothesis. *Journal of the Royal Statistical Society: Series B (Methodological)*, *35*(2), 285-305. https://doi.org/10.1111/j.2517-6161.1973.tb00959.x

Dickey, J. M. (1976). Approximate posterior distributions. *Journal of the American Statistical Association*, *71*(355), 680-689. https://doi.org/10.2307/2285601

Donkin, C., Heathcote, A., & Brown, S. (2009). Is the linear ballistic accumulator model really the simplest model of choice response times: A Bayesian model complexity analysis. Ninth International Conference on Cognitive Modeling—ICCM2009, Manchester,

Dorfman, H. M., & Gershman, S. J. (2019). Controllability governs the balance between Pavlovian and instrumental action selection. *Nature Communications*, *10*(1), 5826. https://doi.org/10.1038/s41467-019-13737-7

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook*

of nonlinear filtering, *12*(656-704), 3.

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, *21*(2), 553-565. https://doi.org/10.1093/bib/bbz016

Evans, N. J. (2019). Assessing the practical differences between model selection methods in inferences about choice response time tasks. *Psychonomic Bulletin & Review*, *26*(4), 1070-1098. https://doi.org/10.3758/s13423-018-01563-9

Evans, N. J., Hawkins, G. E., & Brown, S. D. (2020). The role of passing time in decision-making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 316-326. https://doi.org/10.1037/xlm0000725

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Fong, E., & Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, *107*(2), 489-496. https://doi.org/10.1093/biomet/asz077

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, *26*(4), 1099-1121. https://doi.org/10.3758/s13423-018-1554-2

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*, 641-666. https://doi.org/10.1146/annurev-psych-122414-033645

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics New York.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1), 70-87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, *34*(1), 220-234. https://doi.org/10.1016/j.neuroimage.2006.08.035

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153-160. https://doi.org/10.1080/01621459.1979.10481632

Gelfand, A. E., & Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(3), 501-514. https://doi.org/10.1111/j.2517-6161.1994.tb01996.x

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997-1016. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis. 3rd edition*. Chapman and Hall/CRC. .

Geng, H., Chen, J., Chuan-Peng, H., Jin, J., Chan, R. C. K., Li, Y., Hu, X., Zhang, R.-Y., & Zhang, L. (2022). Promoting computational psychiatry in China. *Nature Human Behaviour*, *6*(5), 615-617.

https://doi.org/10.1038/s41562-022-01328-4

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1-6. https://doi.org/10.1016/j.jmp.2016.01.006

Gronau, Q. F., & Wagenmakers, E. J. (2019). Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection. *Computational Brain & Behavior*, *2*(1), 1-11. https://doi.org/10.1007/s42113-018-0011-7

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E. J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80-97. https://doi.org/10.1016/j.jmp.2017.09.005

Guitart-Masip, M., Huys, Q. J., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage*, *62*(1), 154-166. https://doi.org/10.1016/j.neuroimage.2012.04.024

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis(7th ed.)*. Pearson Prentice Hall.

Hammersley, J. (2013). *Monte carlo methods*. Springer Science & Business Media.

Heck, D. W. (2019). A caveat on the Savage–Dickey density ratio: The case of computing Bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology*, *72*(2), 316-333. https://doi.org/https://doi.org/10.1111/bmsp.12150

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A Conceptual Introduction to Bayesian Model Averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200-215. https://doi.org/10.1177/2515245919898657

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297-307. https://doi.org/10.1093/biomet/76.2.297

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404-413. https://doi.org/10.1038/nn.4238

Huys, Q. J., Cools, R., Golzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Computational Biology*, *7*(4), e1002028. https://doi.org/10.1371/journal.pcbi.1002028

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, *80*(2), 519-530. https://doi.org/10.1016/j.neuron.2013.09.009

Ikink, I., Engelmann, J. B., van den Bos, W., Roelofs, K., & Figner, B. (2019). Time ambiguity during intertemporal decision-making is aversive, impacting choice and neural value coding. *Neuroimage*, *185*, 236-244. https://doi.org/10.1016/j.neuroimage.2018.10.008

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795. https://doi.org/10.1080/01621459.1995.10476572

Kennedy, L., Simpson, D., & Gelman, A. (2019). The Experiment is just as Important as the Likelihood in Understanding the Prior: a Cautionary Note on Robust Cognitive Modeling. *Computational Brain & Behavior*,

2(3-4), 210-217. https://doi.org/10.1007/s42113-019-00051-0

Kording, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Science*, *10*(7), 319-326. https://doi.org/10.1016/j.tics.2006.05.003

Kvålseth, T. O. (1985). Cautionary Note about R^2. *The American Statistician*, *39*(4), 279-285. https://doi.org/10.1080/00031305.1985.10479448

Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLoS Computational Biololgy*, *15*(4), e1006973. https://doi.org/10.1371/journal.pcbi.1006973

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, *14*(10), 1250-1252. https://doi.org/10.1038/nn.2904

Li, J. A., Dong, D., Wei, Z., Liu, Y., Pan, Y., Nori, F., & Zhang, X. (2020). Quantum reinforcement learning during human decision-making. *Nature Human Behaviour*, *4*(3), 294-307. https://doi.org/10.1038/s41562-019-0804-2

Li, Z.-W., & Ma, W. J. (2021). An uncertainty-based model of the effects of fixation on choice. *PLOS Computational Biology*, *17*(8), e1009190. https://doi.org/10.1371/journal.pcbi.1009190

Lotfi, S., Izmailov, P., Benton, G., Goldblum, M., & Wilson, A. G. (2022). *Bayesian Model Selection, the Marginal Likelihood, and Generalization* Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v162/lotfi22a.html

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

McFadden, D. L. (1984). Chapter 24 Econometric analysis of qualitative response models. In *Handbook of Econometrics* (Vol. 2, pp. 1395-1457). Elsevier. https://doi.org/10.1016/S1573-4412(84)02016-X

Menard, S. (2000). Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician*, *54*(1), 17-24. https://doi.org/10.1080/00031305.2000.10474502

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831-860. https://www.jstor.org/stable/24306045

Merlise, C., & Edward, I. G. (2004). Model Uncertainty. *Statistical Science*, *19*(1), 81-94. https://doi.org/10.1214/088342304000000035

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Science*, *16*(1), 72-80. https://doi.org/10.1016/j.tics.2011.11.018

Murphy, K. P. (2023). *Probabilistic machine learning: an introduction*. MIT press.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79-95. https://doi.org/10.3758/BF03210778

Myung, J., & Pitt, M. (2018). Model Comparison in Psychology. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (Vol. VO.5, pp. 1-34). https://doi.org/10.1002/9781119170174.epcn503

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. John Wiley & Sons.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425-433. https://doi.org/10.1016/j.tics.2017.03.011

Pedersen, M. L., Ironside, M., Amemori, K. I., McGrath, C. L., Kang, M. S., Graybiel, A. M., Pizzagalli, D. A., &

Frank, M. J. (2021). Computational phenotyping of brain-behavior dynamics underlying approach-avoidance conflict in major depressive disorder. *PLoS Computational Biololgy*, *17*(5), e1008955. https://doi.org/10.1371/journal.pcbi.1008955

Plummer, M., Stukalov, A., & Denwood, M. (2016). rjags: Bayesian graphical models using MCMC. *R package version*, *4*(6).

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260-281. https://doi.org/10.1016/j.tics.2016.01.007

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593. https://doi.org/10.1126/science.275.5306.1593

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464. https://www.jstor.org/stable/2958889

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333-343. https://doi.org/10.1007/BF02294360

Sivula, T., Magnusson, M., Matamoros, A. A., & Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv*. https://doi.org/10.48550/arXiv.2001.00980

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583-639. https://doi.org/10.1111/1467-9868.00353

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 485-493. http://www.jstor.org/stable/24774528

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, *16*(7), 966-973. https://doi.org/10.1038/nn.3413

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2014). Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision*, *1*(3), 161-183. https://doi.org/10.1037/dec0000005

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2016). Bayes factors for reinforcement-learning models of the Iowa gambling task. *Decision*, *3*(2), 115. https://doi.org/10.1037/dec0000040

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*(4), 1004-1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Stone, M. (1977). An asymptotic equivalence of choice of model by cross‐validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B*, *39*(1), 44-47. https://doi.org/10.1111/j.2517-6161.1977.tb01603.x

Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-theory Methods*, *7*(1), 13-26. https://doi.org/10.1080/03610927808827599

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., Cheng, K., & Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron*, *74*(6), 1125-1137. https://doi.org/10.1016/j.neuron.2012.04.030

Swart, J. C., Frobose, M. I., Cook, J. L., Geurts, D. E., Frank, M. J., Cools, R., & den Ouden, H. E. (2017).

Catecholaminergic challenge uncovers distinct Pavlovian and instrumental mechanisms of motivated (in)action. *Elife*, *6*. https://doi.org/10.7554/eLife.22169

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., & Willemsen, J. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1-26. https://doi.org/10.1038/s43586-021-00017-2

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44-62. https://doi.org/10.1037/a0021765

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In *The Oxford handbook of computational and mathematical psychology.* (pp. 300-319). Oxford University Press.

Vehtari, A. (2022). *Cross-validation FAQ.* https://avehtari.github.io/modelselection/CV-FAQ.html

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413-1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Simpson, D. P., Yao, Y., & Gelman, A. (2019). Limitations of "Limitations of Bayesian Leave-one-out Cross-Validation for Model Selection". *Computational Brain & Behavior*, *2*(1), 22-27. https://doi.org/10.1007/s42113-018-0020-6

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *The Journal of Machine Learning Research*, *17*(1), 3581-3618. http://jmlr.org/papers/v17/14-540.html

Verstynen, T., & Kording, K. P. (2023). Overfitting to 'predict' suicidal ideation. *Nature Human Behaviour*(5), 680-681. https://doi.org/10.1038/s41562-023-01560-6

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228-243. https://doi.org/10.1037/a0027127

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review 14*(5), 779-804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192-196. https://doi.org/10.3758/BF03206482

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158-189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, *11*(12). http://jmlr.org/papers/v11/watanabe10a.html

Westbrook, A., van den Bosch, R., Määttä, J., Hofmans, L., Papadopetraki, D., Cools, R., & Frank, M. J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, *367*(6484), 1362-1366. https://doi.org/10.1126/science.aaz5891

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, *9*(1), 60-62. http://www.jstor.org/stable/2957648

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*, *8*. https://doi.org/10.7554/eLife.49547

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, *92*(4), 937-950. https://doi.org/10.1093/biomet/92.4.937

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917-1007. https://doi.org/10.1214/17-BA1091

Zhang, L., Lengersdorff, L., Mikus, N., Glascher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, *15*(6), 695-707. https://doi.org/10.1093/scan/nsaa089

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95-112. https://doi.org/10.1016/j.jeconom.2015.02.006